Méthodes de Monte Carlo pour l'inférence statistique

Pierre Gloaguen

Cours ENSTA, Avril 2020

Table des matières

Préambule

Ces notes ont pour but de présenter certaines méthodes de simulations dans un but d'applications en statistiques. Nous présenterons d'abord les méthodes pour la simulation de variables aléatoires à partir d'un ordinateur. Nous parlerons ensuite de méthode de Monte Carlo pour l'intégration. Plus précisément, nous verrons comment les méthodes de simulation permettent de faire du calcul approché d'intégrales. Enfin, nous parlerons de l'utilisation des outils de simulation de loi pour l'inférence statistique, et spécifiquement dans le cadre Bayésien.

Ces notes n'ont rien d'original, elles ne font que reprendre, de manière souvent moins exhaustive, des cours existants.

J'ai en effet été puiser dans les cours de différents confrères, en premier lieu dans le cours d'Arnaud Guyader (Sorbonne Université) [3], mais également de Bernard Delyon (Université Rennes I) [2], Sylvain Rubenthaler (Université Nice Sophia Antipolis) [6], Julien Stoehr (Paris-Dauphine) et Sylvain le Corff (Télécom Paris). Je remercie ces auteurs pour le libre accès à leurs notes, qui facilite grandement la diffusion du savoir. Les notes que j'écris sont évidemment libre d'accès et de diffusion en ce sens.

De même la série d'exercices proposées en TD est souvent un échantillon des exercices déjà proposés dans ces références.

1 Introduction

La description d'un phénomène par un modèle probabiliste procure un avantage majeur pour la prise de décision : il permet de quantifier l'incertitude associée au modèle. Ainsi, on pourra construire des outils d'aide à la décision tel que "d'après le modèle posé, la probabilité que tel événement arrive est de ...". Ainsi, être capable de calculer une telle probabilité est un enjeu majeur pour un modèle utile. De manière générale, la grande majorité des décisions statistiques se base sur le calcul d'une intégrale. Comme exemple immédiat, on pensera à la procédure du test statistique. Dans cette procédure où l'on doit décider entre deux hypothèses H_0 et H_1, on décidera sur la base du calcul d'une probabilité (donc d'une intégrale, dans le cas où une fonction de densité existe) sous H_0 (la fameuse probabilité critique).

D'un autre côté, un enjeu majeur en statistiques est, étant donnés un modèle ayant des paramètres inconnus, et un jeu de données (supposé) issu de ce modèle, de définir une procédure d'inférence (un *estimateur*) permettant de "retrouver" les paramètres inconnus.

Plusieurs grandes méthodes existent dans ce cadre :

- Méthode des moments : Les paramètres inconnus sont exprimés sous forme d'une espérance, qui est approchée par les données.
- Maximum de vraisemblance : La loi des observations est vue comme une fonction des paramètres, la vraisemblance. On cherche les paramètres qui maximisent cette vraisemblance (appliquée aux données).
- Inférence Bayésienne : Les paramètres sont supposés être des variables aléatoires suivant une loi (donnée du modèle). L'objectif de l'inférence Bayésienne est de déterminer la loi des paramètres *sachant les données*.

Un point essentiel de ces trois méthodes qui justifie ce cours est le besoin récurrent d'évaluer des espérances :

- Méthode des moments : Par définition, on a besoin de connaître les moments d'une loi et de les lier aux paramètres.
- Maximum de vraisemblance : Dans énormément de modèle, le calcul direct de la vraisemblance n'est pas faisable. On passera alors par des algorithmes intermédiaires, type Espérance-Maximisation, pour maximiser la vraisemblance. Ces algorithmes demandent d'être capable d'évaluer des espérances.
- Inférence Bayésienne : Une fois obtenue la loi des paramètres sachant les observations, on on sera intéressé par des caractéristiques de cette loi (son espérance, la probabilité d'excéder un certain seuil, etc ...).

Encore une fois, toutes ces espérances peuvent s'exprimer sous la forme d'intégrales.

1.0.0.1 Point clé et objectif du cours

Le point clé de cette introduction est de vous faire sentir qu'on a constamment besoin, en statistiques, d'évaluer des intégrales. Une classe de méthodes empiriques et génériques pour ce faire est l'ensemble des méthodes de Monte Carlo. L'objectif de ce cours est de vous présenter ces méthodes, leur fondement théorique et leur mise en pratique.

2 Méthodes de Monte Carlo

2.1 Exemple introductif

Soit φ une fonction sur $\mathbb{R}$ et $a < b$ deux réels.

Supposons que l'on souhaite calculer une intégrale du type

$$I = \int_a^b g(x)\mathrm{d}x$$

On peut remarquer que

$$\begin{aligned} I &= \int_{\mathbb{R}} \overbrace{(b-a)g(x)}^{:=\varphi(x)} \frac{\mathbf{1}_{a\leq x\leq b}}{b-a}\mathrm{d}x \\ &= \mathbb{E}[\varphi(U)], \text{ où } U \sim \mathcal{U}[a,b] \end{aligned}$$

De manière générale, si on veut calculer une intégrale

$$I = \int_{\mathbb{R}^d} \varphi(x)f(x)\mathrm{d}x$$

où f est une fonction positive, telle que $\int_{\mathbb{R}^d} f(x)\mathrm{d}x = 1$, alors on a que

$$I = \mathbb{E}[\varphi(X)] \tag{1}$$

où X est une variable aléatoire de densité f.

La loi des grands nombres fournit alors une manière naturelle d'estimer ce type d'intégrale.

2.2 Résultats asymptotiques

Proposition 2.1. *Loi forte des grands nombres* Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées, et une fonction φ définie sur le support de X_1, telles que $\mathbb{E}[|\varphi(X_1)|] < \infty$, alors :

$$\frac{1}{n}\sum_{k=1}^{n} \varphi(X_k) \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}[X_1]$$

Ainsi, pour approcher I comme dans (1), il suffit de simuler un échantillon $X_1, \ldots, X_n$ selon la densité f, on pose alors l'estimateur :

$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n} \varphi(X_k)$$

Ainsi, l'estimateur est clairement sans biais, et de plus, de par la loi des grands nombres, il est consistant.

Le théorème central limite permettra d'obtenir un intervalle de confiance asymptotique pour l'estimateur.

Proposition 2.2. *Théorème central limite* Avec les notations et les hypothèses de 2.1, avec l'hypothèse supplémentaire que $\mathbb{E}[\varphi(X)^2] < \infty$, alors

$$\sqrt{n}\left(\hat{I}_n - I\right) \xrightarrow[n\to\infty]{Loi} \mathcal{N}(0,\sigma^2)$$

où $\sigma^2 = \mathbb{V}[\varphi(X)]$.

Ainsi, en notant z_α le quantile d'ordre $\alpha \in]0,1[$ de la loi $\mathcal{N}(0,1)$, si on définit l'intervalle aléatoire

$$J_n = \left[\hat{I}_n - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}; \hat{I}_n + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right]$$

Alors,

$$\mathbb{P}(J_n \ni I) \underset{n\to\infty}{\longrightarrow} 1-\alpha$$

J_n est donc un intervalle de confiance asymptotique au niveau 1 - α pour la valeur de I.

En pratique cependant, cet intervalle n'est pas calculable quand σ^2 ne l'est pas.

On dispose cependant d'un estimateur consistant de σ^2 donné par

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\varphi(X_k) - \hat{I}_n\right)^2$$

On peut utiliser alors le lemme de Slutsky.

Proposition 2.3. *Lemme de Slutsky* Soient $(Y_n)_{n\in\mathbb{N}}$ et $(Z_n)_{n\in\mathbb{N}}$ deux suites de variables aléatoires. S'il existe une variable aléatoire Y telle que $(Y_n)_{n\in\mathbb{N}}$ converge en loi vers Y, et une constante c telle que $(Z_n)_{n\in\mathbb{N}}$ converge en probabilité vers c, alors

$$Z_n Y_n \underset{n\to\infty}{\overset{Loi}{\longrightarrow}} cY.$$

Ainsi, par continuité de la fonction $\frac{1}{\sqrt{x}}$, $\frac{1}{\sqrt{\hat{\sigma}_n^2}}$ est un estimateur consistant de $\frac{1}{\sigma}$ et on a

Proposition 2.4.

$$\frac{\sqrt{n}}{\sqrt{\hat{\sigma}_n^2}}\left(\hat{I}_n - I\right) \underset{n\to\infty}{\overset{Loi}{\longrightarrow}} \mathcal{N}(0,1)$$

L'intervalle aléatoire

$$J_n = \left[\hat{I}_n - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}; \hat{I}_n + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\right]$$

donc un intervalle de confiance asymptotique au niveau 1 - α pour la valeur de I.

On finira ce rappel des propriétés par la delta-méthode, pratique quand on a accès à un estimateur d'une fonction de la quantité cible.

Proposition 2.5. *Méthode delta.* Pour toute fonction g dérivable telle que $g'(I) \neq 0$, alors

$$\sqrt{n}\left(g(\hat{I}_n) - g(I)\right) \underset{n\to\infty}{\overset{Loi}{\longrightarrow}} \mathcal{N}(0, (g'(I))^2\sigma^2)$$

2.3 Comparaison avec l'intégration numérique

L'objectif présenté ici est de calculer, en dimension d, une intégrale :

$$\int_{\mathbb{R}^d} \varphi(x)\mathrm{d}x$$

Cette intégrale pourrait très bien se calculer par méthodes numériques (en découpe $\mathbb{R}^d$ en cubes de côtés h, en on considère φ sur ce cube.

Pour une fonction φ de classe C^s, l'erreur est de l'ordre $\frac{1}{n^{\frac{s}{d}}}$.

Pour les méthodes de Monte Carlo, l'écart type de l'erreur est de l'ordre $\frac{1}{n^{\frac{1}{2}}}$, indépendamment de la dimension et de la régularité de φ.

Ainsi, ces méthodes deviennent vite avantageuses quand d est grand.

2.4 Réduction de variance

Il existe de multiples méthodes pour réduire la variance d'un estimateur Monte Carlo.

Pour le lecteur intéressé, on mentionnera, sans les décrire :
- Les variables antithétiques ;
- Les variables de contrôle ;
- Les méthodes de stratification.

Dans cette section, on discutera d'une autre méthode, générique, et très utile, l'échantillonnage préférentiel.

2.4.1 Échantillonnage préférentiel

On cherche à estimer une intégrale du type :

$$I = \int_\Omega \varphi(x) f(x) \mathrm{d}x = \mathbb{E}_X[\varphi(X)]$$

où $\Omega \subset \mathbb{R}^d$, et f est une densité de probabilité sur Ω (on suppose, quitte à renormaliser, que $f(x) = 0$ pour $x \notin \Omega$) et X la variable aléatoire correspondante. Soit g une densité de probabilité telle que $x \in \Omega \Rightarrow g(x) > 0$ et Y la variable aléatoire correspondante, alors il est clair que :

$$I = \int_\Omega \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x = \mathbb{E}\left[\varphi(Y)\frac{f(Y)}{g(Y)}\right]$$

Comme estimateur de I, on peut ainsi proposer l'estimateur :

$$\hat{I}_n^{IS} = \frac{1}{n}\sum_{i=1}^n \varphi(Y_i)\frac{f(Y_i)}{g(Y_i)}$$

où $Y_1, \ldots, Y_n$ est un échantillon i.i.d. de variables aléatoires sur $\mathbb{R}^d$ de densité g.

La variable aléatoire $W(Y_i) = \frac{f(Y_i}{g(Y_i)}$ est appelée poids d'importance de Y_i. Cette appellation sera sans doute plus claire plus bas. On peut voir immédiatement que l'estimateur d'échantillonnage préférentiel reste sans biais.

Intéressons nous à sa variance :

$$\begin{aligned}
\mathbb{V}[\hat{I}_n^{IS}] &= \frac{1}{n}\left(\mathbb{E}_Y\left[\left(\varphi(Y)\frac{f(Y)}{g(Y)}\right)^2\right] - I^2\right) \\
&= \frac{1}{n}\int_{\mathbb{R}^d}\left(\frac{(\varphi(y)f(y))^2}{g(y)^2} - I^2\right) g(y)\mathrm{d}y \\
&= \frac{1}{n}\int_{\mathbb{R}^d}\frac{(\varphi(y)f(y))^2 - I^2 g(y)^2}{g(y)}\mathrm{d}y \\
&= \frac{1}{n}\int_{\mathbb{R}^d}\frac{(\varphi(y)f(y) - Ig(y))^2}{g(y)} + 2I\varphi(y)f(y) - 2I^2 g(y)\mathrm{d}y \\
&= \frac{1}{n}\left\{\int_{\mathbb{R}^d}\frac{(\varphi(y)f(y) - Ig(y))^2}{g(y)}\mathrm{d}y + 2I\left(\int_{\mathbb{R}^d}\phi(y)f(y)\mathrm{d}(y) - I\int_{\mathbb{R}^d} g(y)\mathrm{d}(y)\right)\right\} \\
&= \frac{1}{n}\int_{\mathbb{R}^d}\frac{(\varphi(y)f(y) - Ig(y))^2}{g(y)}\mathrm{d}y
\end{aligned}$$

Donc la variance est nulle quand $g(x) = \frac{\varphi(x)f(x)}{I}$. Ce g optimal n'est pas d'une grande utilité car I est inconnu en pratique. Par contre, l'idée sous jacente doit rester qu'une bonne loi de proposition doit avoir de la masse là où $\varphi \times f$ a de la masse. Il faut donc trouver une densité dont la masse est un compromis entre là où f a de la masse, et là où φ prend des grandes valeurs.

Notons également que si g est très petit là où $\varphi \times f$ est non négligeable, alors la variance sera très grande !

2.4.2 Échantillonnage préférentiel normalisé

Supposons, cas fréquent, que f ne soit connue qu'à une constante près.

C'est à dire que l'on ait accès qu'à une fonction positive (non normalisée) $f^{(u)}$ telle que $\int f^{(u)} = c$ On a lors, $f(x) = f^{(u)}(x)/c$.

Dans ce cas, on peut toujours approcher l'intégrale :

$$I = \mathbb{E}[\varphi(X)] = \int_\Omega \varphi(x) f(x) d(x)$$

par l'estimateur :

$$\hat{I}_n^{IS,u} = \sum_{i=1}^n \varphi(Y_i) \frac{f^{(u)}(Y_i)/g(Y_i)}{\sum_{\ell=1}^n f^{(u)}(Y_\ell)/g(Y_\ell)}$$

Proposition 2.6. $\hat{I}_n^{IS,u} \overset{Proba}{\longrightarrow} I$

Démonstration. On définit :

$$\begin{aligned} w_i &= \frac{f(Y_i)}{g(Y_i)} \\ w_i^{(u)} &= \frac{f^{(u)}(Y_i)}{g(Y_i)} \\ \tilde{w}_i &= \frac{w_i}{\sum_{\ell=1}^n w_\ell} \\ \tilde{w}_i^{(u)} &= \frac{w_i^{(u)}}{\sum_{\ell=1}^n w_\ell^{(u)}} \end{aligned}$$

On remarque, par définition de $f^{(u)}$ que $\tilde{w}_i = \tilde{w}_i^{(u)}$, donc :

$$\hat{I}_n^{IS,u} = \sum_{i=1}^n \tilde{w}_i^{(u)} \varphi(Y_i) = \sum_{i=1}^n \tilde{w}_i \varphi(Y_i) = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{w}_i \varphi(Y_i)}{\frac{1}{n}\sum_{\ell=1}^n \tilde{w}_\ell}$$

Le numérateur converge presque sûrement vers I (il s'agit de l'estimateur IS vu plus haut). Quand au dénominateur, en constatant que $\mathbb{E}[w_i] = 1$, on a qu'il converge presque sûrement vers 1. Ainsi, par la proposition 2.3, le ratio converge en loi vers la constante I (ce qui est équivalent à une convergence en probabilité vers cette constante). □

On peut également caractériser la variance asymptotique de l'estimateur.

Proposition 2.7.

$$\sqrt{n}\left(\hat{I}_n^{IS,u} - I\right) \overset{Loi}{\longrightarrow} \mathcal{N}(0, \mathbb{V}[\hat{I}_1^{IS} - IW(Y_1)])$$

Démonstration. On notera tout d'abord une extension naturelle de la méthode delta (Propriété 2.5). Pour un vecteur $\beta \in \mathbb{R}^d$, un estimateur $\hat{\beta}_n$ de β et une fonction C^1 $h : \mathbb{R}^d \mapsto \mathbb{R}$ telle que le gradient en β ne s'annule pas, alors :

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \overset{Loi}{\longrightarrow} \mathcal{N}(0, \Sigma) \Rightarrow \sqrt{n}\left(h(\hat{\beta}_n) - h(\beta)\right) \overset{Loi}{\longrightarrow} \mathcal{N}(0, \nabla h(\beta)^T \Sigma \nabla h(\beta))$$

où Σ est une matrice de covariance $d \times d$.

Notons $\hat{W}_n = \frac{1}{n}\sum_{i=1}^{n} w_i$. On définit alors $\hat{\beta}_n$, β et Σ comme :

$$\hat{\beta}_n = \begin{pmatrix} \hat{I}_n^{IS} \\ \hat{W}_n \end{pmatrix}$$
$$\beta = \begin{pmatrix} I \\ 1 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} \mathbb{V}[\hat{I}_1^{IS}] & \mathbb{C}\text{ov}[\hat{I}_1^{IS}, \hat{W}_1] \\ \mathbb{C}\text{ov}[\hat{I}_1^{IS}, \hat{W}_1] & \mathbb{V}[\hat{W}_1] \end{pmatrix}.$$

La fonction h est alors la fonction $h(x, y) = x/y$. On a alors :

$$\nabla h(\beta)^T \Sigma \nabla h(\beta)) = \mathbb{V}[\hat{I}_1^{IS} - I\hat{W}_1]$$

□

3 Simulation de variables aléatoires

Dans ce chapitre, nous abordons plusieurs méthodes de simulation de variables aléatoires.

Pour une loi de probabilité donnée (par exemple, la loi normale), le calcul de probabilités permet de calculer des quantités associées, telles que l'espérance, la variance, la fonction de répartition. Cependant, peut on définir une méthode algorithmique permettant de générer un échantillon aléatoire issu de cette loi ? C'est à dire une suite de variable aléatoires $X_1, \ldots, X_n$ telles qu'elles soient mutuellement indépendantes, et distribuées selon cette loi ?

En pratique, il n'existe pas aujourd'hui de méthode générique de simulation *aléatoire* [1] par ordinateur. Cependant, des algorithmes *déterministes* ont été proposés pour *mimer* un comportement de variables aléatoires indépendantes et identiquement distribuées. Ces algorithmes sont appelés *générateurs pseudo-aléatoires*. En pratique, la simulation d'une variable aléatoire de loi quelconque se ramènera de manière "atomique" à la simulation d'une loi uniforme sur l'intervalle $[0,1]$ [2]. On décrira donc dans la section suivante un générateur pseudo aléatoire pour la loi uniforme sur $[0,1]$.

3.1 Générateur à congruences pour la loi $\mathcal{U}[0,1]$

Dans cette section, on présente un algorithme déterministe de simulation pour simuler une loi uniforme sur l'intervalle $[0,1]$. Le but de ce polycopié n'est pas de couvrir le très vaste sujet des générateurs pseudos aléatoires. Le lecteur intéressé pourra se référer au chapitre 3 de [4]. Les notations de cette section sont d'ailleurs issues de cette référence.

Algorithme Une méthode générique pour mimer le comportement d'un échantillon de variables aléatoires $U_1, \ldots, U_n$ indépendantes et identiquement distribuées de loi $\mathcal{U}[0,1]$ la méthode de congruence linéaire.

L'algorithme se base sur 4 données initiales choisies par l'utilisateur :
- Un entier $m > 0$, appelé *module* ;
- Un entier $0 < a < m$ appelé *multiplicateur* ;
- Un entier $0 \leq c < m$ appelé *incrément* ;
- Un entier $0 \leq x_0 < m$ appelé *graine*.

On créera alors une suite de nombres $x_1, \ldots x_n$ en utilisant la relation de récurrence

$$x_k == (ax_{k-1} + c) \text{ modulo } m.$$

On définit enfin les nombres $u_1, \ldots, u_n$ tels que $u_k = \frac{x_k}{m}$, $1 \leq k \leq n$, qui sont, par construction, dans l'intervalle $[0,1]$. La méthode est entièrement résumé par l'Algorithme 1.

Entrée : 5 entiers m, a, c, x_0 et n
Sortie : $u_1, \ldots, u_n$: réalisation pseudo-aléatoire d'un échantillon i.i.d. de loi $\mathcal{U}[0,1]$

1 **pour** $k \leftarrow 1$ **à** n **faire**
2 $\quad x_k \leftarrow (ax_{k-1} + c)$ modulo m;
3 $\quad u_k \leftarrow x_k/m$;
4 **fin**

Algorithme 1 : Générateur à congruences

La séquence $u_1, \ldots, u_n$ ainsi générée est donc à valeurs dans $[0,1]$ [3]. Pour des valeurs de a, c et m bien

1. Au sens commun du terme, c'est à dire *imprévisible* exactement par quiconque
2. Autrement dit, toute quantité aléatoire utilisée sera algorithmiquement obtenue à partir de transformation déterministes de lois uniformes.
3. Plus exactement dans $[0,1[$ telle que définie ici. Si on prend $c = 0$ et m premier, elle est même à valeurs dans $]0,1[$.

choisies, cette séquence se comporte comme la réalisation d'un échantillon de n variables aléatoires $U_1, \ldots, U_n$ indépendantes et identiquement distribuées de loi $\mathcal{U}[0,1]$.

Un générateur à congruences correspond au choix de a, c et m. La graine x_0 correspond au point de départ de l'algorithme. Pour un x_0 fixé, la séquence obtenue en sortie sera *toujours* la même. En pratique, quand un tel algorithme est appelé dans un ordinateur, la graine n'est pas demandée à l'utilisateur, mais obtenue en interne. Une méthode générique est de prendre le nombre de millisecondes (modulo m) écoulé depuis le 1er Janvier 1970.

Choix de a, c et m : Le choix des constantes du générateur est une question délicate.

Un premier point important est que la "loi" de l'échantillon obtenu doit mimer celle de la loi cible. Pour ce faire, on pourra utiliser un test statistique (voir la section 3.4.

Un autre point est que l'échantillon simuler doit mimer l'indépendance. Or, il faut d'ores et déjà remarquer que chaque x_k est dans l'ensemble fini $\{0, \ldots, m-1\}$, ainsi, la suite $(x_n)_{n\geq 0}$ est nécessairement *périodique*.

Un facteur souhaité pour mimer l'aléatoire est que cette période ne soit pas "visible" par l'utilisateur, ainsi on voudra qu'elle soit longue, ce qui implique que m soit grand. De même, pour que la période soit grande, il faut en pratique que a soit grand, et, si possible, relativement premier à m. Un autre aspect pris en compte doit être la rapidité de l'opération "modulo m". Nous n'irons pas plus point concernant ces points, largement discutés dans [4]. Un tableau des valeurs utilisées pour les générateurs les plus connus et disponible dans le chapitre 1 de [2].

Dans la suite du cours, on supposera que l'on dispose d'une méthode valide de simulation de variables uniformes indépendantes. La plupart des langages informatiques en dispose. Dans le logiciel `R`, cette méthode est implémentée dans la fonction `runif`.

3.2 Méthodes de simulation de lois

On s'intéresse désormais à simuler une variable aléatoire quelconque.

3.2.1 Rappels sur la fonction de répartition

Définition 3.1. *Fonction de répartition* Soit X une variable aléatoire à valeurs réelles. Pour tout réel x, on appelle fonction de répartition de X la fonction F_X :

$$\begin{array}{rcl} \mathbb{R} & \mapsto & [0,1] \\ x & \mapsto & F_X(x) = \mathbb{P}(X \leq x) \end{array}$$

Une fonction de répartition F_X est caractérisée par les propriétés suivantes :

1. F_X est partout continue à droite, i.e. pour tout $x \in \mathbb{R}$:

$$\lim_{\substack{h \to 0 \\ >0}} F(x+h) = F(x)$$

2. F_X est croissante.
3. $\lim_{x\to-\infty} F_X(x) = 0, \lim_{x\to+\infty} F_X(x) = 1$

Ainsi, toute fonction F sur $\mathbb{R}$ satisfaisant ces conditions est une fonction de répartition.

Des exemples de fonctions de répartitions sont montrés sur la Figure 1.

Une fonction importante définie à partir de la fonction de répartition est son inverse généralisée. Cette fonction est l'outil de la base de la méthode d'inversion décrite plus bas.

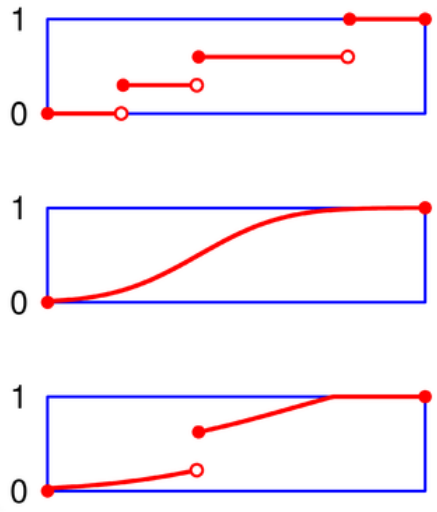


FIGURE 1 – Exemples de fonction de répartition pour une variable aléatoire discrète (haut), continue (centre) ou avec atome (bas). Source *Wikipedia*.

Définition 3.2. *Inverse généralisée* Soit F une fonction de répartition, on appelle inverse généralisée de F, notée, F^{-1} la fonction :

$$\begin{array}{rcl}]0,1[& \mapsto & \mathbb{R} \\ u & \mapsto & F^{-1}(u) = \inf\{z \in \mathbb{R} \text{ tel que } F(z) \geq u\} \end{array}$$

Pour une variable aléatoire X, la fonction F_X^{-1} est également appelée *fonction quantile* de la variable aléatoire X. Dans ce cas, on convient que $F_X^{-1}(0)$ et $F_X^{-1}(1)$ sont la plus petite et la plus grande des valeurs possibles pour X (éventuellement infinies).

Remarque : Dans le cas d'une fonction de répartition F continue et strictement croissante sur $\mathbb{R}$ [4], la fonction F^{-1} est simplement l'inverse de F.

Définition 3.3. *Fonction de répartition empirique*

Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées selon la loi d'une variable aléatoire X.

La fonction de répartition empirique de la variable aléatoire X associée à $X_1, \ldots X_n$, notée F_X^n est la fonction :

$$\begin{array}{rcl} \mathbb{R} & \mapsto & [0,1] \\ x & \mapsto & F_X^n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} \end{array}$$

On vérifie facilement que F_X^n est une fonction de répartition.

3.2.2 Méthode d'inversion

Supposons qu'on connaisse la fonction de répartition de X, F_X, alors on a la propriété suivante (propriété d'inversion) :

Proposition 3.1. *Méthode d'inversion*

Soit F une fonction de répartition. Soit F^{-1} son inverse généralisée. Soit U une variable aléatoire de loi uniforme sur $[0,1]$, alors la variable aléatoire

$$X := F^{-1}(U)$$

4. Associée à une variable aléatoire continue sur $\mathbb{R}$, par exemple

admet F comme fonction de répartition.

Démonstration. On veut montrer que, pour tout $x \in \mathbb{R}$

$$\mathbb{P}(F^{-1}(U) \leq x) = F(x).$$

Montrons tout d'abord que, pour tout $u \in]0,1[$

$$\forall x \in \mathbb{R}, F^{-1}(u) \leq x \Leftrightarrow u \leq F(x).$$

$\Rightarrow$ Soient $u \in]0,1[$ et $x \in \mathbb{R}$ tels que $F^{-1}(u) \leq x$.
Par croissance de F, on a donc :

$$F\left(F^{-1}(u)\right) \leq F(x)$$

Or, en se souvenant que par définition

$$F^{-1}(u) = \inf\left\{z \in \mathbb{R} \text{ tel que } F(z) \geq u\right\},$$

on a donc directement

$$u \leq F\left(F^{-1}(u)\right) \leq F(x)$$

$\Leftarrow$ Soient $u \in]0,1[$ et $x \in \mathbb{R}$ tels que $u \leq F(x)$.
Ainsi, $x \in \{z \in \mathbb{R} \text{ tel que } F(z) \geq u\}$, donc $F^{-1}(u) \leq x$.

On a donc montré, l'équivalence. Il reste à conclure en se servant de la définition d'une loi uniforme :

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

□

Conséquence et intérêt de la Proposition 3.1 : Comme conséquence immédiate de cette proposition, on obtient une méthode de simulation d'un échantillon pour une variable aléatoire X de loi de probabilité F_X :

1. Calculer F_X^{-1}
2. Simuler un échantillon aléatoire i.i.d. $U_1, \ldots, U_n$de loi $\mathcal{U}[0,1]$ [5]
3. Obtenir un échantillon i.i.d. $X_1, \ldots, X_n = F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n)$

3.2.3 Méthode d'acceptation rejet

La méthode d'acceptation rejet permet de simuler selon une densité f, évaluable en tout point, même lorsque la méthode d'inversion ne peut être appliquée.

L'idée est de se servir d'une loi qu'on sait simuler (la loi de proposition), de densité g, et ayant un support couvrant celui de f.

Proposition 3.2. *Méthode d'acceptation-rejet* Soit f et g deux densités sur $\mathbb{R}^d$. On suppose qu'il existe une constante M telle que

$$\forall x \in \mathbb{R} \quad f(x) \leq M(g(x))$$

On note

$$\alpha(x) := \frac{f(x)}{Mg(x)}.$$

Soit $(U_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de densité g. On note T la variable aléatoire (à valeurs dans $\mathbb{N}^*$) :

$$T = \inf\left\{n, \text{ tel que } U_n \leq \alpha(Y_n)\right\}.$$

. Alors, la variable aléatoire $X := Y_T$ (T-ième valeur de la suite $(Y_n)_{n\geq 1}$ a pour densité f).

5. Qu'on est capable de simuler grâce à un algorithme similaire à ceux décrit en 3.1.

Démonstration. Soit un entier $n \leq 1$:

$$\mathbb{P}\left(X \leq x, T = n\right) = \mathbb{P}\left(U_1 > \alpha(Y_1), \ldots, U_{n-1} > \alpha(Y_{n-1}), U_n \leq \alpha(Y_n), Y_n \leq x\right)$$

Par propriété d'indépendance

$$= \mathbb{P}\left(U_n \leq \alpha(Y_n), Y_n \leq x\right) \prod_{i=1}^{n-1} \mathbb{P}\left(U_i > \alpha(Y_i)\right)$$

Par propriété de distribution identique

$$= \mathbb{P}\left(U_n \leq \alpha(Y_n), Y_n \leq x\right) \mathbb{P}\left(U_1 > \alpha(Y_1)\right)^{n-1}$$

Or on a :

$$\begin{aligned}
\mathbb{P}\left(U_1 > \alpha(Y_1)\right) &= \mathbb{E}\left[\mathbf{1}_{U_1 > \alpha(Y_1)}\right] \\
&= \int_{\mathbb{R}} \left(\int_0^1 \mathbf{1}_{u > \frac{f(y)}{Mg(y)}} \mathrm{d}u \right) \times g(y) \mathrm{d}y \\
&= \int_{\mathbb{R}} (1 - \frac{f(y)}{Mg(y)}) g(y) \mathrm{d}y
\end{aligned}$$

Comme f et g sont des densités :

$$\mathbb{P}\left(U_1 > \alpha(Y_1)\right) = 1 - \frac{1}{M}$$

De manière analogue

$$\begin{aligned}
\mathbb{P}\left(U_n \leq \alpha(Y_n), Y_n \leq x\right) &= \mathbb{E}\left[\mathbf{1}_{U_n \leq \alpha(Y_n)} \times \mathbf{1}_{Y_n \leq x}\right] \\
&= \int_{\mathbb{R}} \left(\int_0^1 \mathbf{1}_{u \leq \frac{f(y)}{Mg(y)}} \mathrm{d}u \right) \times \mathbf{1}_{y \leq x} g(y) \mathrm{d}y \\
&= \int_{-\infty}^{x} \frac{f(y)}{M} \mathrm{d}y \\
&= \frac{F(x)}{M},
\end{aligned}$$

où $F(x)$ est la fonction de répartition associée à f. En résumé :

$$\mathbb{P}\left(X \leq x, T = n\right) = \frac{F(x)}{M} \left(1 - \frac{1}{M}\right)^{n-1}$$

Donc on conclut en remarquant que

$$\mathbb{P}(X \leq x) = \sum_{n=1}^{\infty} \mathbb{P}\left(X \leq x, T = n\right) = F(x)$$

□

Remarque sur la loi du temps d'attente : De la preuve, on peut déduire que

$$\mathbb{P}\left(T = n\right) = \lim_{x \to \infty} \mathbb{P}\left(X \leq x, T = n\right) = \lim_{x \to \infty} \frac{F(x)}{M} \left(1 - \frac{1}{M}\right)^{n-1} = \frac{1}{M} \left(1 - \frac{1}{M}\right)^{n-1}$$

Donc, la loi du temps d'attente avant d'obtenir une réalisation de F est une loi géométrique de paramètre $\frac{1}{M}$.

L'espérance du nombre d'essais avant le premier succès est donc M.

Pour un g donné, l'algorithme est optimal en complexité quand M est minimal.

Cas de lois discrètes La proposition 3.3 permet de simuler pour des variables aléatoires à densité. Un analogue existe pour les variables aléatoires discrètes.

Proposition 3.3. *Méthode d'acceptation-rejet, cas discret* Soit X une variable aléatoire discrète sur l'ensemble $\{1, \dots, K\}$ et Y une variable aléatoire discrète sur l'ensemble $\{1, \dots, K'\}$ (avec $K' \geq K$). On suppose qu'il existe une constante M telle que

$$\forall k \in \{1, \dots, K'\} \quad \mathbb{P}(X = k) \leq M\mathbb{P}(Y = k)$$

On note

$$\alpha(k) := \frac{\mathbb{P}(X = k)}{M\mathbb{P}(Y = k)}.$$

Soit $(U_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0, 1]$. Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de même loi que Y. On note T la variable aléatoire (à valeurs dans $\mathbb{N}^*$) :

$$T = \inf \{n, \text{ tel que } U_n \leq \alpha(Y_n)\}.$$

Alors, la variable aléatoire Y_T (T-ième valeur de la suite $(Y_n)_{n\geq 1}$ a la même loi que X.

Démonstration. La preuve est l'analogue pour les lois discrètes de la précédente. □

3.2.4 Utilisation de transformation usuelles

On peut bien évidemment utiliser la théorie des probabilités pour simuler différentes lois, notamment les règles sur la stabilité des lois :

- Si $c > 0$ et $d \in \mathbb{R}$ $U \sim \mathcal{U}[a, b] \Rightarrow cU + d \sim \mathcal{U}[ac + d, bc + d]$
- Si $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Etc...

3.3 Vecteurs aléatoires

Supposons qu'on veuille simuler un vecteur aléatoire (X, Y) où X et Y sont deux variables aléatoires réelles (cela s'étend évidemment à un vecteur de taille n).

Un cas simple est quand X et Y sont indépendantes, alors, on se ramène au cas univarié, et le couple aléatoire est donné par deux réalisations indépendantes.

Dans les autres cas, certaines méthodes peuvent s'avérer utiles.

3.3.1 Conditionnement

Un cas utile en pratique est le cas où l'on sait simuler facilement X, et l'on sait facilement simuler Y sachant X.

On utilise alors la décomposition de la loi jointe

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(Y \leq y | X \leq x) \times \mathbb{P}(X \leq x)$$

Exemple Supposons qu'on veuille simuler selon la densité jointe :

$$f_{X,Y}(x, y) = x\mathrm{e}^{-xy}\mathbf{1}_{0\leq x\leq 1}\mathbf{1}_{y>0}$$

Alors, la densité marginale de X est donnée par :

$$f_X(x) = \int_0^\infty x\mathrm{e}^{-xy}\mathbf{1}_{0\leq x\leq 1}\mathrm{d}y = \mathbf{1}_{0\leq x\leq 1}$$

Ainsi, X suit une loi uniforme sur $[0,1]$. x étant fixé, la loi de $Y|X$ est donnée par une loi exponentielle de paramètre x. On peut donc facilement simuler (X,Y) en simulant selon une uniforme, puis une exponentielle de paramètre donné par cette uniforme.

3.3.2 Changement de variables

Pour les couples de variables aléatoires à densité, on peut utiliser la propriété du changement de variables.

Proposition 3.4. Soit un couple de variables aléatoires (U,V) de densité $f_{U,V}(u,v)$ définie sur $E_{UV} \subset \mathbb{R}^2$ et un couple de variables aléatoires (X,Y) à valeurs dans $E_{XY} \subset \mathbb{R}^2$. Supposons qu'il existe une application ϕ, C^1, inversible, et d'inverse C_1, tel que $(X,Y) = \phi(U,V)$, alors la densité jointe de (X,Y) est donnée par :

$$f_{X,Y}(x,y) = f_{U,V}(\phi^{-1}(x,y))|\det J_{\phi^{-1}}(x,y)|$$

où J_ϕ désigne la matrice jacobienne d'une application $\phi(u,v)$:

$$J_\phi(u,v) = \begin{pmatrix} \frac{\delta\phi_1}{\delta u}(u,v) & \frac{\delta\phi_1}{\delta v}(u,v) \\ \frac{\delta\phi_2}{\delta u}(u,v) & \frac{\delta\phi_2}{\delta v}(u,v) \end{pmatrix}$$

3.3.3 Simulation d'un vecteur Gaussien

En pratique, un type de vecteur aléatoire important est le vecteur Gaussien. En dimension d, un vecteur Gaussien $X = (X_1,\ldots,X_d)'$ est un vecteur dont toutes les combinaisons linéaires des composantes sont des lois Gaussiennes sur $\mathbb{R}$.

Il est entièrement caractérisé par le vecteur $\mathbb{E}[X] = (E[X_1],\ldots,\mathbb{E}[X_d])'$ et sa matrice de variance (covariance) Σ :

$$\Sigma_{i,j} = \mathbb{C}\mathrm{ov}[X_i, X_j]$$

Il se trouve qu'un vecteur Gaussien est stable par transformation affine.

Ainsi, si X est un vecteur Gaussien de paramètres $\mu := \mathbb{E}[X]$ et de variance Σ, alors pour un vecteur b de taille d et une matrice $d \times d$, alors

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A').$$

Cette propriété implique qu'on peut simuler tout vecteur Gaussien à partir de d variables aléatoires Gaussiennes indépendantes centrées et réduites. Pour une matrice de variance Σ voulu, il suffit de trouver R tel que $RR' = \Sigma$. Ce calcul est toujours possible car Σ est une matrice semi-définie positive.

3.4 Complément : Test statistique d'adéquation à une loi

Dans ce chapitre, nous avons défini des algorithmes de simulation d'échantillons aléatoires.

Correctement évaluer si un algorithme reproduit un comportement aléatoire est un sujet très vaste en informatique, qui ne sera pas abordé ici (encore une fois, voir [4], ou encore, [1], chapitre 7 sur la théorie de la complexité).

D'un point de vue statistique, la question d'évaluer si un échantillon observé est la réalisation d'un échantillon aléatoire de loi connue est également un problème d'intérêt.

Ce problème peut être traité par différents tests statistiques. Nous en présenterons ici 2 :
- Le test du χ^2 d'adéquation (pour l'adéquation a une loi discrète) ;
- Le test de Kolmogorov-Smirnoff, pour l'adéquation a une loi continue

3.4.1 Test du χ^2 d'adéquation

Le test du χ^2 d'adéquation à une loi permet de tester si un échantillon i.i.d. de v.a. discrètes $X_1, \ldots, X_n$, à valeurs dans l'ensemble $\{1, \ldots, K\}$ est distribuée selon une loi multinomiale de paramètre $p^* = (p_1, \ldots, p_K)$.

On définit $N_k = \sum_{i=1}^n \mathbf{1}_{X_i=k}$ et la variable aléatoire

$$T_n = \sum_{k=1}^{K} \frac{(N_k - np_k)^2}{np_k}$$

3.4.1.1 Propriété : Quand $n \mapsto \infty$, T_n converge en loi vers un $\chi^2(K-1)$.

Pour le test :
H_0 La loi de X_1 est p^* ;
H_1 La loi de X_1 n'est pas p^*,
On utilisera T_n comme statistique de test.

Pour un risque de première espèce $\alpha \in]0,1[$, on aura alors la procédure de test suivante,

Pour un échantillon observé $x_1, \ldots, x_n$, on définit $n_k = \sum_{i=1}^n \mathbf{1}_{x_i=k}$. On calcule $t_n = \sum_{k=1}^K \frac{(n_k - np_k)^2}{np_k}$ On rejette alors H_0 au risque α si $t_n > z_{1-\alpha}$ ou $z_{1-\alpha}$ est le quantile d'ordre 1 - α d'une loi du $\chi^2(n-1)$.

3.4.2 Test de Kolmogorov-Smirnoff

Le test de Kolmogorov-Smirnoff permet de tester l'adéquation échantillon i.i.d. de v.a. continues $X_1, \ldots, X_n$ à une loi continue de fonction de répartition F^*.

On définit la variable aléatoire

$$T_n = \sqrt{n} \max_{i=1\ldots,n} |F_n(X_i) - F(X_i)|$$

Propriété : Quand $n \mapsto \infty$, T_n converge en loi vers la variable aléatoire T de fonction de répartition F_T :

$$F_T(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - 2\sum_{k=1}(-1)^{k+1} \exp\left(-2k^2x^2\right) & \text{sinon} \end{cases}$$

Cette loi est appelée loi de Kolmogorov Smirnoff.

Pour un risque de première espèce $\alpha \in]0,1[$, on aura alors la procédure de test suivante,

Pour le test :
H_0 La loi de X_1 est F^* ;
H_1 La loi de X_1 n'est pas F^*,
On utilisera T_n comme statistique de test.

Pour un échantillon observé $x_1, \ldots, x_n$, on calcule la fonction de répartition empirique associée F_n et le t_n correspondant.

On rejette alors H_0 au risque α si $1 - F_T(t_n) > 1 - \alpha$

Ces deux tests peuvent être utiles pour vérifier qu'une méthode de simulation fonctionne.

Nous allons maintenant nous intéresser aux méthodes de simulation proprement dites.

4 Inférence bayésienne

4.1 Rappel sur le maximum de vraisemblance

En statistique paramétrique, on suppose qu'un ensemble d'observations $\mathbf{X}$ est la réalisation d'une variable aléatoire dont la loi dépend d'un ensemble de paramètres θ inconnu et à valeurs dans un espace Θ. L'inférence statistique consiste en la définition d'un estimateur de θ.

Un estimateur générique commun est l'estimateur du maximum de vraisemblance.

Le modèle statistique posé permettant d'écrire la loi de $\mathbf{X}$ quand on connaît θ, que l'on note $L(\mathbf{X}, \theta)$. On choisit comme estimateur le paramètre $\hat{\theta}$ *le plus vraisemblable*, c'est à dire celui qui maximise (en θ) la fonction $L(\mathbf{X}, \theta)$.

L'estimateur du maximum de vraisemblance pour $\mathbf{X}$ est donné par $\hat{\theta} = \text{argmax}_\theta L(\theta|\mathbf{X}) = \frac{\sum_{i=1}^n X_i}{n}$.

Cet estimateur **est une variable aléatoire**. Sa loi dépend du modèle, mais asymptotiquement, un thérème central limite nous assure que sa distribution devient celle d'une loi Normale (dont l'expression de la variance est connue, au moins en théorie).

Dans ce contexte, le paramètre θ est donc une quantité fixe inconnue. Toute la connaissance sur sa valeur vient des données.

4.2 Inférence bayésienne

4.2.1 Connaissance a priori et définition du posterior

Dans le contexte de l'inférence bayésienne, on supposera que le paramètre θ est lui même aléatoire. On modélisera alors sa loi sous la forme d'une distribution. Cette distribution est indépendante des données et s'appelle la distribution *a priori*. Elle reflète la connaissance (et l'incertitude) que l'on a sur le paramètre. La loi a priori sur θ sera notée $\pi(\theta)$.

L'objectif de l'inférence bayésienne est d'actualiser cette connaissance (et son incertitude) grâce au données. La quantité d'intérêt, dans ce contexte est alors la loi de $\theta|\mathbf{X}$, quand appelle loi a posteriori (ou posterior). Cette quantité sera notée $\pi(\theta|\mathbf{X})$.

La formule de Bayes sur le conditionnement permet de lier cette loi a posteriori à la loi a priori et à la vraisemblance du modèle :

$$\pi(\theta|\mathbf{X}) = \frac{\pi(\theta)L(\mathbf{X}|\theta)}{\int_\Theta \pi(u)L(\mathbf{X}|u)}du$$

On remarque que le dénominateur ne dépend pas de θ, il s'agit d'une constante de normalisation. On écrira souvent cette relation

$$\pi(\theta|X) \propto \pi(\theta)L(\mathbf{X}|\theta)$$

Ce sont les caractéristiques de cette loi (ses quantiles, ses moments) que l'on cible dans le contexte de l'inférence bayésienne.

L'objectif de l'inférence bayésienne est donc la détermination de $\pi(\theta|\mathbf{X})$.

4.2.2 Choix du prior

Pour un nombre de données limité, la **forme du prior** a un impact sur la forme du posterior.

La forme du prior peut être choisie en fonction du *savoir expert* (littérature existante, expériences passées).

ATTENTION : Le support du posterior sera toujours inclu dans le support du prior.

Si le prior charge tout le support de manière égale, on dit qu'il est **non informatif**.

4.2.3 Prior impropre

Si le support de θ est sur $\mathbb{R}$, un prior non informatif est une "uniforme sur $\mathbb{R}$". Ceci n'est pas cependant pas une loi de probabilité !

On peut cependant noter abusivement $\pi(\theta) \propto 1$. Dans ce cas, si $\frac{L(\mathbf{X})|\theta)}{\int_\Theta L(\mathbf{X}|\theta)\mathrm{d}\theta}$ définit une loi de probabilité en θ, alors le posterior $\pi(\theta|\mathbf{X})$ est bien défini. Le prior est alors dit **impropre**.

4.2.4 Estimateurs bayésiens

Les estimateurs bayésiens sont des quantités liées à la loi à posteriori.

On mentionnera :

- Le maximum a posteriori (MAP), correspondant à à la valeur de θ maximisant $\pi(\theta|\mathbf{X})$.
- L'espérance a posteriori
$$\mathbb{E}[\theta|\mathbf{X}] = \int_\theta \pi(\theta|\mathbf{X})\theta d\theta.$$
- Intervalles de crédibilités : Pour toute région $\mathcal{R} \subset \Theta$, on peut quantifier :
$$\mathbb{P}(\theta \in \mathcal{R}|\mathbf{X}) = \int_\mathcal{R} \pi(\theta|\mathbf{X})\mathrm{d}\theta$$
Pour $\alpha \in]0,1[$, une région de crédibilité de niveau $1-\alpha$ est une région $\mathcal{R} \subset \Theta$ telle que
$$\mathbb{P}(\theta \in \mathcal{R}|\mathbf{X} = \mathbf{x}) = 1 - \alpha$$
Cet intervalle n'est pas asymptotique, mais **dépend du prior**.

4.2.5 Détermination du posterior

Il existe deux cas différents en inférence bayésienne :

- Soit la loi a posteriori est dans une famille connue (loi normale, loi beta, etc,...), alors l'inférence est directe, et tous les estimateurs bayésiens peuvent être obtenus facilement.
- Soit la loi a posteriori n'appartient pas à une famille de loi connue. Dans ce cas, il faudra obtenir les quantités d'intérêt par méthode de Monte Carlo. Pour cela, il faudra souvent être capable d'obtenir un échantillon i.i.d. selon la loi a posteriori. Les méthodes vues jusqu'alors pourront être utilisées. On verra qu'elles ne suffiront pas toujours, et que d'autres méthodes, les méthodes de Monte Carlo par chaîne de Markov, aideront à s'en sortir.

Une manière astucieuse de se retrouver dans le cas 1 est d'utiliser les propriétés de conjugaisons de certaines lois. On parlera alors de priors conjugués au modèle.

Les deux sections suivantes décrivent chacune un exemple illustratif de ces cas.

4.3 Cas conjugué : modèle beta-binomial

4.3.1 Expérience et question

On suppose qu'on dispose d'une pièce, et l'on souhaite déterminée si elle est équilibrée. Pour cela, on effectue n tirages indépendant de pile ou face.

4.3.2 Modélisation

On note $\mathbf{x} = (x_1, \ldots, x_n)$ le résultat du lancer (0 si *face*, 1 si *pile*). On suppose que ces nombres sont les réalisations d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ où les $X_1, \ldots, X_n$ sont indépendantes et identiquements distribuées de loi $\mathcal{B}ern(\theta)$ où $\theta \in]0,1[$ est la probabilité d'obtenir pile.

Donc, la loi jointe de $\mathbf{X} = (X_1, \ldots, X_n)$ (donc la vraisemblance pour θ) est donnée par :

$$L(\mathbf{X}|\theta) = \prod_{k=1}^{n} \mathbb{P}_\theta(X = X_k) = \theta^{\sum_{k=1}^n X_k} (1-\theta)^{n-\sum_{k=1}^n X_k}$$

où $X \sim \mathcal{B}ern(\theta)$.

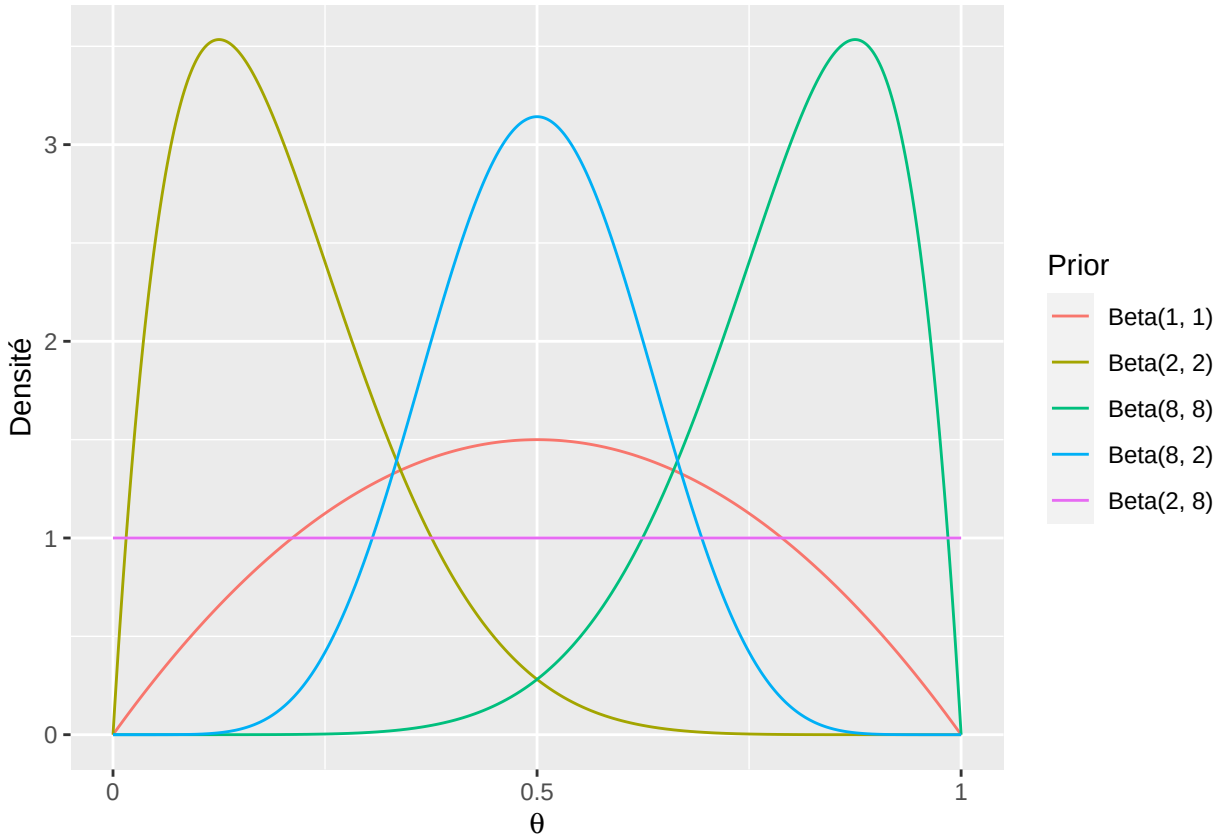
4.3.3 Prior

Pour l'inférence bayésienne, on pose comme *a priori* que $\theta \sim \mathcal{B}eta(a, b)$. Cette loi est censée illustrée notre connaissance indépendante des données sur θ. Le premier point trivial est que l'on sait que θ est entre 0 et 1, donc on a choisi une loi ayant ce support.

Ensuite, le choix des paramètres a et b déterminera notre a priori sur al pièce :

- Le cas $a = b = 1$, correspond à une loi uniforme. Cela traduira un a priori non informatif sur θ, chaque valeur entre $]0, 1[$ nous semble également vraisemblable.
- Le cas $a = b$ avec des valeurs supérieures à 1 traduira un a priori où la pièce est équilibrée. De grandes valeurs de a et b traduiront une plus grande certitude.
- Le cas $a > b$, traduira un a priori où la pièce est déséquilibrée en faveur de pile.
- Le cas $a < b$, traduira un a priori où la pièce est déséquilibrée en faveur de face.

La figure ci dessous illustre ces différents a priori :



L'expression analytique du prior est donc donnée par :

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}\,du}\mathbf{1}_{0<\theta<1} \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1}$$

4.3.4 Loi a posteriori

On cherche la loi de $\theta|\mathbf{X}$.

On a directement que :

$$\begin{aligned}\pi(\theta|\mathbf{X}) &\propto L(\mathbf{X}|\theta)\pi(\theta)\\ &\propto \theta^{\sum_{k=1}^n X_k}(1-\theta)^{n-\sum_{k=1}^n X_k}\theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{0<\theta<1}\\ &\propto \theta^{a+\sum_{k=1}^n X_k-1}(1-\theta)^{b+n-\sum_{k=1}^n X_k-1}\mathbf{1}_{0<\theta<1}\end{aligned}$$

On reconnaît que $\pi(\theta|\mathbf{X})$ est la densité d'une loi

$$\theta|\mathbf{X} \sim \beta\left(a+\underbrace{\sum_{k=1}^n X_k}_{\text{Nb. piles}},\ b+\underbrace{n-\sum_{k=1}^n X_k}_{\text{Nb. faces}}\right)$$

Le fait que la *loi a posteriori* soit dans la même famille que la loi *a priori* est une propriété de conjugaison du modèle binomial avec le prior de loi beta. Ce prior est dit conjugué.

4.3.5 Estimateurs bayésiens

— **Maximum a posteriori (MAP)**

On peut montrer que, pour $a+b+n>2$ et $a+\sum_{k=1}^n x_k \geq 1$

$$MAP(\theta|\mathbf{X}) = \frac{a+\sum_{k=1}^n X_k - 1}{a+b+n-2}$$

On remarque que pour $a=b=1$ (prior uniforme), il s'agit du maximum de vraisemblance, et que pour tout couple (a,b), cette quantité converge vers le maximum de vraisemblance quand n grandit.

— **Espérance a posteriori**

Par propriété de la loi β, on :

$$\mathbb{E}[\theta|\mathbf{X}] \overset{\text{loi } \beta}{=} \frac{a+\sum_{k=1}^n X_k}{a+b+n} = \underbrace{\frac{n}{a+b+n}}_{\text{Poids données}} \times \overbrace{\frac{\sum_{k=1}^n X_k}{n}}^{\text{Max. de vrais.}} + \underbrace{\frac{a+b}{a+b+n}}_{\text{Poids prior}} \times \overbrace{\frac{a}{a+b}}^{\mathbb{E}\text{ du prior}}$$

Encore une fois, la décomposition illustre le poids des données et le poids du prior. On remarque que pour n suffisament grand, tous les priors seront équivalents.

— **Régions de crédibilité**

Toute région de crédibilité peut facilement être obtenue à l'aide de la fonction quantile de la loi β, qui est implémentée dans tout logiciel de statistiques.

4.4 Posterior de loi inconnue : modèle de régression probit :

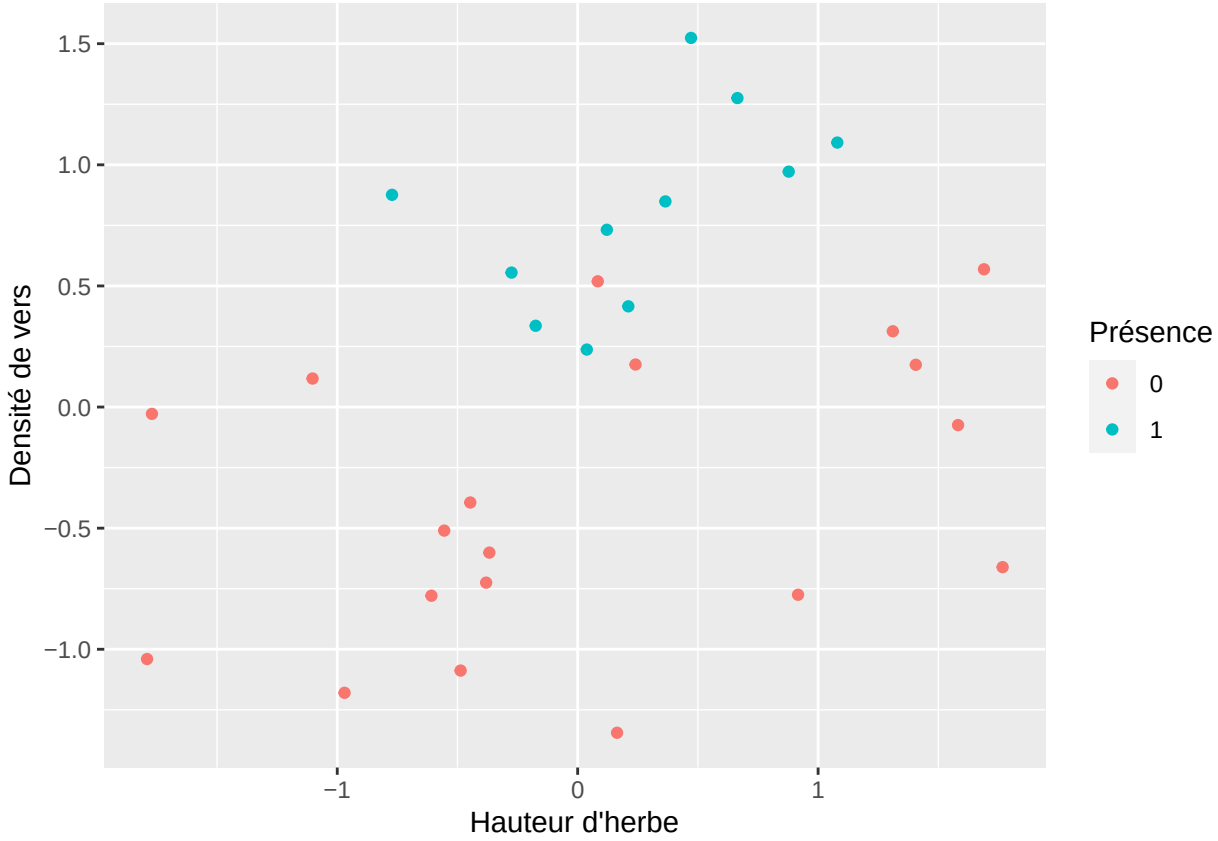
4.4.1 Prédiction de présence d'oiseaux

Une étude consiste en l'observation de la présence ou non de la linotte mélodieuse sur différents sites échantillonnés.

Sur ces différents sites sont mesurées différentes caractéristiques :

— Le nombre de vers moyens sur une surface au sol de $1m^2$. (Covariable 1)
— La hauteur d'herbe moyenne sur une surface au sol de $1m^2$. (Covariable 2)
— On calcule cette hauteur d'herbe au carré. (Covariable 3).

On peut représenter la présence ou non d'oiseaux en fonctions des caractéristiques du site :



4.4.2 Notations et modèle de régression probit

On note $y_1, \ldots, y_n$ les observations de présence (1 si on observe un oiseau, 0 sinon) sur les sites 1 à n.

On note

$$\mathbf{x}_k = (\overset{\text{Nb. vers}}{x_{k,1}}, \overset{\text{Haut. herbe}}{x_{k,2}}, \overset{\text{Haut. herbe}^2}{x_{k,3}})^T$$

le vecteur des covariables sur le k-ème site ($1 \leq k \leq n$).

On pose le modèle suivant :

$Y_k \sim \mathcal{B}ern(p_k)$ où

$$p_k = \phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) = \phi(\mathbf{x}_k^T \theta),$$

où

— ϕ est la fonction de répartition d'une $\mathcal{N}(0,1)$, i.e.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u$$

— $\theta = \{\beta_0, \beta_1, \beta_2, \beta_3\}$ est le vecteur des paramètres à estimer.

4.4.3 Vraisemblance

Pour un vecteur d'observations $\mathbf{Y} = Y_{1:k}$, la vraisemblance

$$L(\mathbf{Y}|\theta) = \prod_{k=1}^{n} \underbrace{\phi(\mathbf{x}_k^T\theta)^{Y_k}}_{\text{Proba. présence}} \times \underbrace{(1-\phi(\mathbf{x}_k^T\theta))^{1-Y_k}}_{\text{Proba. absence}}$$

4.4.4 Prior sur θ

Comme a priori sur θ, on choisit une normale avec une grande variance$\theta \overset{\text{prior}}{\sim} \mathcal{N}(0, 4I)$, donc

$$\pi(\theta) = \frac{1}{\sqrt{2\pi \times 4}^4} \mathrm{e}^{-\frac{1}{8}\theta^T\theta}$$

où I est la matrice Identité (ici 4×4)

4.4.5 Posterior

Le posterior est donc donné par :

$$\pi(\theta|\mathbf{Y}) \propto \pi(\theta)L(\mathbf{Y}|\theta) \propto \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{Y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-Y_k}$$

Cette densité n'est pas standard. Ainsi, on ne connaît pas ces caractéristiques associées et intéressantes (quantiles, espérance, variance). Ces différentes quantités peuvent cependant être approchées par méthode de Monte Carlo, si tant est qu'on soit capable de simuler selon cette loi.

4.4.6 Simulation d'échantillons a posteriori par acceptation rejet

Notre objectif est de simuler selon le posterior défini ci dessus. On va pour ce faire procéder par méthode d'acceptation rejet.

On remarque immédiatement qu'on ne peut pas utiliser l'acceptation rejet classique, car $\pi(\theta|\mathbf{Y})$ n'est connu qu'à une cnstante près !

Cependant, si on note

$$\tilde{\pi}(\theta|\mathbf{Y}) = \frac{1}{64\pi^2} \mathrm{e}^{-\frac{1}{8}\theta^T\theta} \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{Y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-Y_k},$$

on peut utiliser la propriété vue en TD, qui dit qu'il suffit de connaître $\tilde{\pi}$ pour simuler selon π à partir d'acceptation rejet.

— **Choix de la densité de proposition**

Comme densité de proposition, on peut par exemple prend pour g la densité correspondant au prior ($g(\theta) = \pi(\theta)$). On remarque que dans ce cas

$$\frac{\tilde{\pi}(\theta|\mathbf{Y})}{g(\theta)} = \frac{\pi(\theta)L(\mathbf{Y}|\theta)}{\pi(\theta)} = \prod_{k=1}^{n} \phi(\mathbf{x}_k^T\theta)^{Y_k}(1-\phi(\mathbf{x}_k^T\theta))^{1-Y_k} \leq 1 =: M.$$

On a donc un majorant uniforme en θ et l'acceptation rejet suivant permet de tirer selon $\pi(\theta|\mathbf{Y})$:

1. On tire $\theta_{cand} \sim \mathcal{N}(0, 4I)$
2. On tire (independamment) $U \sim \mathcal{U}[0, 1]$
3. Si $U < \frac{L(y_{1:n}|\theta)}{M}$, on accepte θ_{cand}
4. Sinon on recommence

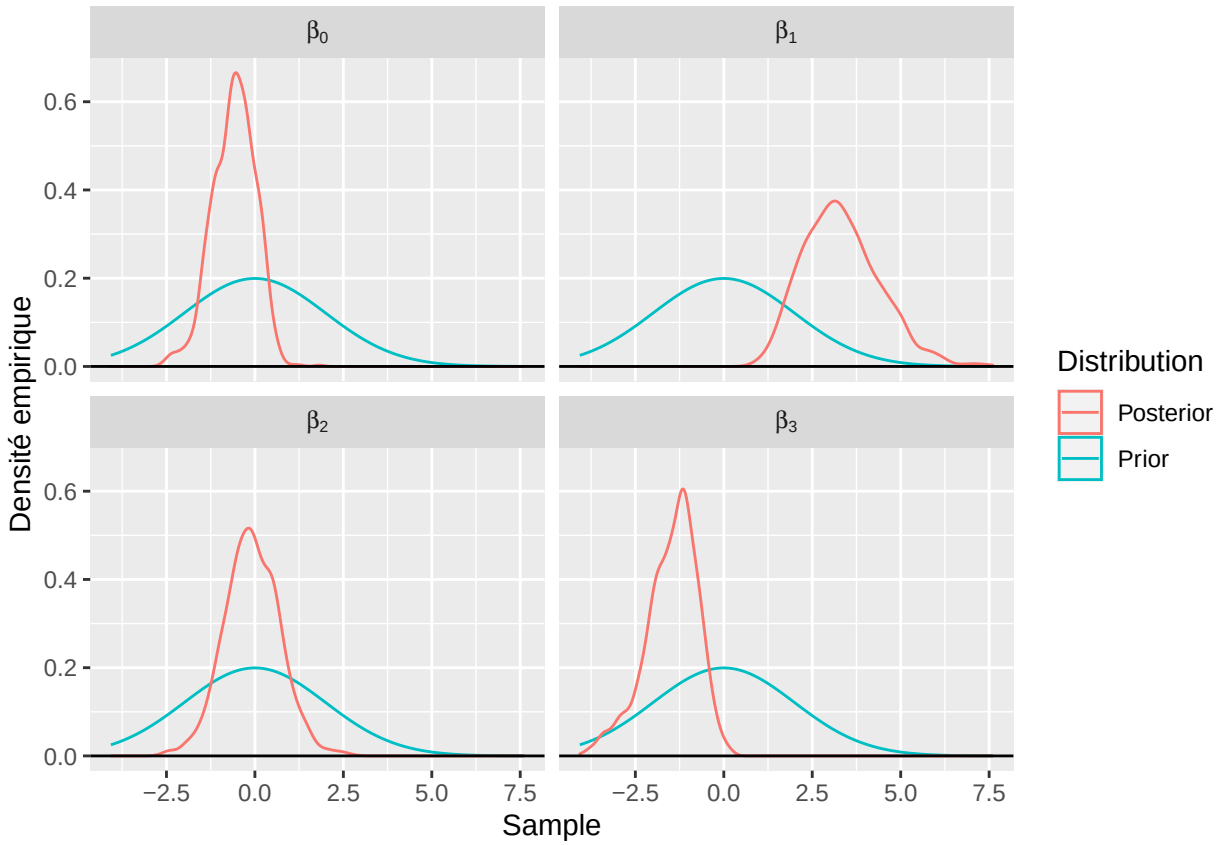
```
## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
```

```
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".

## Warning: UNRELIABLE VALUE: Future ('<none>') unexpectedly generated random
## numbers without specifying argument 'seed'. There is a risk that those random
## numbers are not statistically sound and the overall results might be invalid.
## To fix this, specify 'seed=TRUE'. This ensures that proper, parallel-safe random
## numbers are produced via the L'Ecuyer-CMRG method. To disable this check, use
## 'seed=NULL', or set option 'future.rng.onMisuse' to "ignore".
```

4.4.7 Echantillon du posterior, loi a posteriori marginales et estimateurs bayésiens

— **Lois marginales**

On effectue un tirage de taille $M = 1000$. On peut représenter la densité empirique de cet échantillon

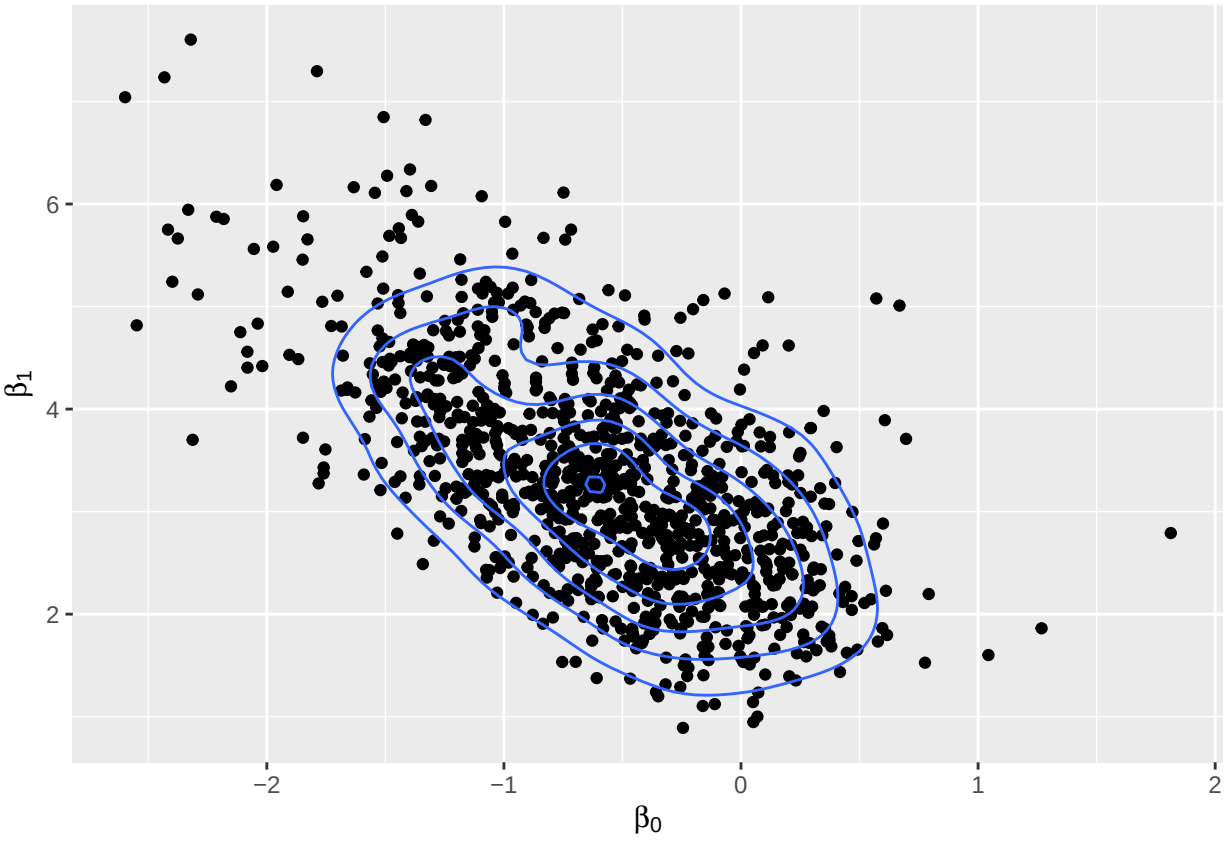


On peut remarquer que les densités du posterior sont différentes du prior.

— **Loi jointe**

Il peut être intéressé de regarder la loi jointe obtenue par simulation. On pourra notamment remarquer qu'un prior indépendant sur les composantes n'entraîne pas une indépendance dans le posterior. Par exemple, on représente ici la loi du couple $(\beta_0, \beta_1|\mathbf{Y})$:

Paramètre	Esperance a posteriori.	Quantile 2.5\%	Quantile 97.5\%
beta[0]	-0.591	-1.847507	0.4395118
beta[1]	3.320	1.534582	5.6908270
beta[2]	-0.082	-1.616159	1.4627349
beta[3]	-1.482	-3.247421	-0.2811471



On voit qu'il existe une correlation linéaire négative entre ces deux variables aléatoires, qui étaient, a priori, supposées indépendantes.

— **Espérance a posteriori et intervalles de crédibilité**

Comme on sait simuler selon la loi cible, les espérances peuvent être approchées par méthode de Monte Carlo classique. On obtient ici une estimation del'espérance **a posteriori** ainsi qu'un intervalle de confiance a posterirori

4.5 Au delà de l'acceptation rejet

Dans le cas précédent, l'espérance du temps d'attente avant une acceptation est donnée par

$$\frac{M}{\int L(\mathbf{Y}|\theta)\pi(\theta)\mathrm{d}\theta}$$

Mécaniquement, cette quantité augmente quand n augmente, et l'acceptation rejet dvient prohibitif.

En pratique, l'inférence Bayésienne utilisera d'autres algorithmes de simulations de loi : les algorithmes de Monte Carlo par chaîne de Markov.

5 Méthodes de Monte Carlo par chaîne de Markov

Il n'est pas toujours possible de simuler (efficacement) un échantillon i.i.d. selon une densité cible f.

Cependant, il existe des outils permettant de construire des chaînes de Markov dont la distribution asymptotique sera donnée par f. La chaîne de Markov ainsi construite permettra l'approximation d'espérance par rapport à f grâce à une version "Chaîne de Markov" de la loi des grands nombres.

Le but de ce chapitre est de présenter l'algorithme de Metropolis-Hastings, méthode très générique pour construire une chaîne de Markov adaptée.

Cet algorithme est très générique et est définit aussi bien que pour des lois discrètes, que pour des lois à densité.

On introduira ici l'algorithme (et la preuve de son fonctionnement), dans le cas discret. La cas continu est présenté bien plus rigoureusement dans [5].

5.1 Rappel sur les chaînes de Markov

On se place dans un ensemble $\mathcal{K}$ fini (typiquement l'ensemble $\{1, \ldots, K\}$).

Définition 5.1. *Chaîne de Markov* Soit X_0 une variable aléatoire sur $\mathcal{K}$ de loi π_0. On dit que la suite de variables aléatoires $(X_n)_{n\geq 0}$ à valeurs $\mathcal{K}$ dans est une chaîne de Markov si pour tout $n \geq 1$ est pour tout suite $(k_0, \ldots, k_n)$ d'éléments de $\mathcal{K}$, on a :

$$\mathbb{P}\left(X_n = k_n | X_0 = k_0, \ldots, X_{n-1} = k_{n-1}\right) = \mathbb{P}(X_n = k_n | X_{n-1} = k_{n-1})$$

On dit que cette chaîne de Markov est *homogène* si, pour i et j dans $\mathcal{K}$: $\mathbb{P}(X_n = j | X_{n-1} = i) = \mathbb{P}(X_1 = j | X_0 = i) = P_{ij}$ La matrice $P = (P_{ij})$ est alors appelée matrice de transition de la chaîne de Markov. P_{ij} est la probabilité de transition de i vers j. Dans la suite on se placera dans le cas homogène. Une chaîne de Markov homogène est entièrement caractérisée par π_0 et P.

On remarque immédiatement que $0 \leq P_{ij} \leq 1$ et $\sum_j P_{ij} = 1$

5.1.0.1 Propriétés On rappelle quelques propriétés utiles. On note π_n, pour $n \geq 0$ la loi de l'état X_n, c'est à dire le vecteur ligne

$$\pi_n = (\pi_{n,1} = \mathbb{P}(X_n = 1), \ldots, \pi_{n,K} = \mathbb{P}(X_n = K)).$$

On a alors :

- $\mathbb{P}(X_1 = j) = \sum_{i=1}^{k} \mathbb{P}(X_0 = i) \times \mathbb{P}(X_1 = j | X_0 = i) = \sum_{i=1}^{k} \pi_{0,i} P_{ij}$ Cette relation est résumée par l'équation $\pi_1 = \pi_0 P$
- On peut montrer par récurrence que

  $$P_{ij}^{(n)} := \mathbb{P}(X_n = j | X_0 = i) = (P^n)_{ij}$$

  où P^n est la puissance n-ième de la matrice P.
- On a ainsi :

  $$\pi_n = \pi_0 P^n$$

Définition 5.2. *Mesure invariante pour P* Soit π un vecteur (ligne) de probabilité (une mesure de probabilité sur $\mathcal{K}$). On dit que π est invariante pour la chaîne de Markov de transition P si :

$$\pi P = \pi$$

Comme conséquence immédiate, on a que si π_0 est une mesure invariante pour P, alors, pour tout n, $\pi_n = \pi_0$. Dans ce cas, les variables aléatoires $X_0, \ldots, X_n$ sont identiquement distribuées (mais pas indépendantes !).

Définition 5.3. *Irréductibilité de P* On dit qu'une chaîne de Markov homogène sur $\mathcal{K}$, de transition P est irréductible si

$$\forall i, j \in \mathcal{K} \times \mathcal{K},\ \exists\ n \text{ tel que } P_{i,j}^{(n)} > 0$$

Autrement dit, pour deux états de la chaîne, il est possible d'accéder de l'un à l'autre en un temps fini.

Définition 5.4. Chaîne apériodique Soit $(X_n)_{n\geq 1}$ une chaîne de Markov homogène sur $\mathcal{K}$. Pour $k \in \mathcal{K}$, on appelle *période* de l'état k, notée $d(k)$, le P.G.C.D. de tous les entiers n tels que $P_{kk}^{(n)} > 0$ (avec la convention $pgcd(\emptyset) = +\infty$) :

$$d(j) = pgcd\left\{n \geq 1, P_{kk}^{(n)} > 0\right\}$$

Une chaîne est dite apériodique si pour tout k dans $\mathcal{K}$, $d(k) = 1$. Pour une chaîne irréductible, une condition suffisante pour être apériodique est qu'il existe un $k \in \mathcal{K}$ tel que $P_{kk} > 0$.

Proposition 5.1. *Théorème ergodique* Soit $(X_n)_{n\geq 0}$ une chaîne de Markov de loi initiale π_0 et de matrice de transition P. On suppose que cette chaîne est irréductible et apériodique. Alors :

1. Cette chaîne de Markov admet une unique mesure de probabilité invariante π.
2. $X_n \xrightarrow{loi} X$ où X est une v.a. de loi π.
3. Pour toute fonction φ intégrable par rapport à π, on a :

$$\frac{1}{n+1}\sum_{i=0}^{n} \varphi(X_i) \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}_\pi[\varphi(X)].$$

4. Si $\varphi(X)$ admet un moment d'ordre supérieur à 2, on a

$$\sqrt{n}\left(\frac{1}{n+1}\sum_{i=0}^{n} \varphi(X_i) - \mathbb{E}_\pi[\varphi(X)]\right) \xrightarrow[n\to+\infty]{Loi} \mathcal{N}(0, \sigma^2)$$

Un analogue de cette propriété reste vrai quand la chaîne de Markov est à valeurs dans un ensemble continu (typiquement, $\mathbb{R}^d$). Si le résultat reste moralement le même, il demande un formalisme plus conséquent. On pourra se référer à [5] pour une présentation rigoureuse.

5.1.0.2 Remarque sur le Théorème Central Limite

Contrairement au TCL dans le cas i.i.d., la variance σ^2 n'est absolument pas triviale (il ne s'agit pas de $\mathbb{V}[\varphi(X)]$!), et n'est pas nécessairement facile à estimer. En effet, les variables aléatoires dans l'estimateur Monte Carlo ne sont plus indépendantes, on ne peut donc plus considérer la variance de l'estimateur comme la somme des variances. Ainsi, obtenir un estimateur de la variance (et donc un intervalle de confiance asymptotique) n'est pas nécessairement évident.

5.1.0.3 Conséquence et intérêt pratique de la Proposition 5.1

Le théorème ergodique pour les chaînes de Markov stipule qu'il n'est pas nécessaire de savoir simuler un échantillon i.i.d. pour obtenir une approximation de l'espérance selon une loi π (ou une densité f, ce qui nous intéressera plus souvent en pratique). En effet, si l'on est capable de construire une chaîne de Markov irréductible de mesure de probabilité invariante π (ou f), alors, en simulant selon cette chaîne de Markov suffisamment longtemps, on pourra approcher toute espérance relativement à π. De plus, le point 2. de la proposition nous assure qu'au bout d'un certain temps, les X_n obtenus pourront être considérés comme de loi π (mais pas indépendants !). Encore faut il être capable de construire une chaîne de Markov (une matrice P) irréductible, de loi invariante donnée par π.

C'est le sens de l'algorithme de Metropolis Hastings.

5.2 Algorithme de Metropolis-Hastings

5.2.1 Définitions

Définition 5.5. *Réversibilité* Soit $\pi = (\pi_1, \dots, \pi_K)$ une mesure de probabilité sur $\mathcal{K}$ et $(X_n)_{n\geq 0}$ une chaîne de Markov homogène de matrice de transition P. On dit que π est réversible pour P si elle vérifie la condition d'équilibre :

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K},\ \pi_i \times P_{ij} = \pi_j \times P_{ji}$$

Proposition 5.2. *Réversibilité* $\Rightarrow$ *Invariance* Si une mesure de probabilité π est réversible pour une chaîne de Markov de transition P, alors, π est une mesure de probabilité invariante pour P.

Démonstration. Soit π une mesure de probabilité réversible pour P. On a tout de suite que

$$\begin{aligned}
\forall j \in \mathcal{K}\ \ (\pi P)_j &= \sum_{i=1}^{K} \pi_i P_{ij} \\
&= \sum_{i=1}^{K} \pi_j P_{ji} && \text{par réversibilité} \\
&= \pi_j && \text{par propriété de } P \\
\Rightarrow \pi P &= \pi
\end{aligned}$$

□

Cette propriété va nous permettre de construire une chaîne de Markov satisfaisant les hypothèses de la Proposition 5.1.

5.2.2 Algorithme dans le cas discret

Proposition 5.3. *Algorithme de Metropolis Hastings* Soit π une mesure de probabilité sur $\mathcal{K}$ selon laquelle on aimerait simuler (on supposera que $\pi_k > 0$ pour tout k). Soit π_0 une mesure de probabilité sur $\mathcal{K}$ telle que $\pi_k > 0 \Rightarrow \pi_{0,k} > 0$ et Q une matrice stochastique $\mathcal{K} \times \mathcal{K}$ satisfaisant la condition suivante :

$$\forall (i,j) \in \mathcal{K} \times \mathcal{K}, Q_{ij} > 0 \Leftrightarrow Q_{ji} > 0$$

On considère la suite de variables aléatoires $(X_n)_{n\geq 0}$ construite de la manière suivante :

1. On simule X_0 selon π_0.
2. Pour $n \geq 1$:
   (a) On tire Y_n selon la loi $Q_{X_{n-1}\bullet}$ (la ligne de Q donnée par X_{n-1}).
   (b) On tire une loi uniforme U indépendante de Y_n.
   (c) On calcule la quantité
   $$\alpha(X_{n-1}, Y_n) = \min\left(1, \frac{\pi_{Y_n} Q_{Y_n X_{n-1}}}{\pi_{X_{n-1}} Q_{X_{n-1} Y_n}}\right)$$
   (d) On pose :
   $$X_n = \begin{cases} Y_n & \text{si } U \leq \alpha(X_{n-1}, Y_n) \\ X_{n-1} & \text{sinon} \end{cases}$$

alors, $(X_n)_{n\geq 1}$ est une chaîne de Markov de transition P où

$$P_{ij} = \begin{cases} Q_{ij}\alpha(i,j) & \text{si } i \neq j \\ 1 - \sum_{j\neq i} P_{ij} & \text{sinon} \end{cases}$$

De plus π est invariante pour P.

Démonstration. La loi de $X_n|X_{0:(n-1)}$ ne dépendant que de X_{n-1}, la propriété de Markov est évidente par construction. Pour $i \neq j$, on a

$$\begin{aligned}\mathbb{P}(X_n = j|X_{n-1} = i) &= \mathbb{P}\left(Y_n = j, U \leq \alpha(X_{n-1}, Y_n)|X_{n-1} = i\right) \\ &= \mathbb{P}\left(Y_n = j, U \leq \alpha(i,j)|X_{n-1} = i\right) \\ &= \mathbb{P}\left(Y_n = j|X_{n-1} = x\right)\mathbb{P}\left(U \leq \alpha(i,j)|X_{n-1} = i\right) \\ &= Q_{ij}\alpha(i,j)\end{aligned}$$

Ce qui prouve la première partie. De plus, pour $i \neq j$, on a :

$$\begin{aligned}\pi_i P_{ij} &= \pi_i Q_{ij}\alpha(i,j) \\ &= \pi_i Q_{ij}\min\left(1, \frac{\pi_j Q_{j,i}}{\pi_i Q_{i,j}}\right) \\ &= \min\left(\pi_i Q_{ij}, \pi_j Q_{j,i}\right) \\ &= \pi_j Q_{j,i}\min\left(\frac{\pi_i Q_{ij}}{\pi_j Q_{j,i}}, 1\right) \\ &= \pi_j Q_{j,i}\alpha(j,i) \\ &= \pi_j P_{j,i}\end{aligned}$$

Donc, π est réversible pour P, donc par la proposition 5.2, π est invariante. On remarquera que si Q est irréductible et apériodique, P l'est aussi, et ainsi, les conditions de la proposition 5.1 sont satisfaites. □

5.2.3 Algorithme dans le cas continu

Supposons qu'on veuille simuler dans $\mathbb{R}^d$ selon une densité f, éventuellement connue à une constante près, c'est à dire que

$$\forall x \in \mathbb{R}^d,\ f(x) = Cf^{(u)}(x) = \frac{f^{(u)}(x)}{\int_{\mathbb{R}^d} f^{(u)}(z)\mathrm{d}z}$$

On remplace alors la matrice de transition par un *noyau de transition* sur $\mathbb{R}^d$, à savoir une fonction

$$\begin{array}{rccc} q: & \mathbb{R}^d \times \mathbb{R}^d & \mapsto & \mathbb{R}_+ \\ & (x,y) & \mapsto & q(x,y) \geq 0 \end{array}$$

telle que $\int_{\mathbb{R}^d} q(x,y)\mathrm{d}y = 1$ (typiquement, la loi d'une marche aléatoire centrée en x.

Alors, si on sait simuler, pour x fixé, selon q, et qu'on a $q(x,y) > 0 \Leftrightarrow q(y,x) > 0$, alors, l'algorithme de Metropolis reste valide en remplaçant π par $f^{(u)}$ et Q par q. On notera que dans le ratio on a pas besoin de la constante de normalisation car

$$\frac{f^{(u)}(y)}{f^{(u)}(x)} = \frac{f(y)}{f(x)}$$

5.2.3.1 Conséquence et intérêt de l'algorithme de Metropolis Hastings L'algorithme de Metropolis Hastings est très puissant, car il permet, sous des conditions faibles, de simuler une chaîne de Markov satisfaisant le théorème ergodique. Il s'agit d'un des algorithmes les plus utilisés en pratique. Un de ces défauts est sa nécessité d'avoir une étape d'acceptation rejet. Ainsi, si la probabilité d'acceptation est faible, la chaîne simulée pourra rester "coincée" dans un état, et la variance de l'estimateur Monte Carlo sera très grande.

5.3 Échantillonneur de Gibbs

On considère une vecteur aléatoire en dimension d $X = (X^{(1)}, \ldots, X^{(d)})$. Dans le cas où la loi f (ou la loi π) est une densité dans en grande dimension ($\mathbb{R}^d$ ou $\mathcal{K}^d$), l'espace à visiter est typiquement grand, et le ratio d'acceptation de l'algorithme de Metropolis Hastings sera souvent assez faible.

Un algorithme très utilisé dans ce cas est l'échantillonneur de Gibbs.

Cet algorithme suppose que l'on sait simuler selon **toutes les lois conditionnelles de** X.

Plus formellement, si on note $X^{-(\ell)} = (X^{(1)}, \dots, X^{(\ell-1)}, X^{(\ell+1)}, X^{(d)})$, on est capable de simuler facilement la variable aléatoire $X^{(\ell)}|X^{(-\ell)}$, selon la loi $\pi^{(-\ell)}$. Dans ce cas, l'idée est la suivante :

1. Prendre $X_0 = (X_0^{(1)}, \dots, X_0^{(d)})$ tiré selon une loi initiale.
2. Pour $k \geq 1$:
   (a) Tirer ℓ uniformément dans $\{1, \dots, d\}$ Actualiser $X_k^{(1)}$ en simulant selon la loi de $X_k^{(1)}|X_{k-1}^{(2)}, \dots, X_{k-1}^{(d)}$;
   (b) Simuler Y selon la loi $X^{(\ell)}|\{X^{(-\ell)} = X_{k-1}^{(-\ell)}\}$
   (c) Poser $X_k = (X_{k-1}^{(1)}, \dots, X_{k-1}^{(\ell-1)}, Y, X_{k-1}^{(\ell+1)}, X_{k-1}^{(d)})$

On peut montrer que l'échantillonneur de Gibbs est équivalent à un algorithme de Metropolis Hastings où la quantité α est toujours égale à 1, c'est à dire un Metropolis Hastings où ne rejette jamais le candidat.

Encore une fois, cet algorithme fonctionne est utile dès que la simulation des lois conditionnelles est faisable. Si les lois conditionnelles induisent une matrice de transition (ou un noyau) de Markov irréductible et apériodique, alors le théorème ergodique s'applique.

Références

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006.

[2] B. Delyon. Simulation et modélisation. Polycopié de cours, https://perso.univ-rennes1.fr/bernard.delyon/simu.pdf.

[3] A. Guyader. Méthodes monte-carlo. Polycopié de cours, http://www.lpsm.paris/pageperso/guyader/files/teaching/MonteCarlo/MonteCarlo.pdf.

[4] D. E. Knuth. *Art of computer programming, volume 2 : Seminumerical algorithms*. Addison-Wesley Professional, 1997.

[5] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[6] S. Rubenthaler. Méthodes de monte-carlo. Polycopié de cours, https://math.unice.fr/~rubentha/enseignement/poly-cours-monte-carlo-m1-im.pdf.